Machine learning classification of surgical pathology reports and chunk recognition for information extraction noise reduction

Napolitano, G., Marshall, A., Hamilton, P., & Gavin, A. T. (2016). Machine learning classification of surgical pathology reports and chunk recognition for information extraction noise reduction. *Artificial Intelligence in Medicine*, *70*, 77-83. https://doi.org/10.1016/j.artmed.2016.06.001

Published in:
Artificial Intelligence in Medicine

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

doi:10.1016/j.artmed.2016.06.001

Machine learning classification of surgical pathology reports and chunk recognition for information extraction noise reduction

Giulio Napolitano
g.napolitano@imbie.uni-bonn.de

Institut für Medizinische Biometrie, Informatik und Epidemiologie (IMBIE)
Universität Bonn
Haus 325/11/1.OG/Raum 620
Sigmund-Freud-Straße 25
53105 Bonn
Germany

Tel: +49 (228) 287 15414
Fax: +49 (228) 287 15032

Adele Marshall
Queen's University Belfast, School of Mathematics and Physics
University Road
Belfast BT7 1NN, United Kingdom

Peter Hamilton
Queen's University Belfast, School of Medicine, Dentistry and Biomedical Sciences
97 Lisburn Road
Belfast BT9 7BL, United Kingdom

Anna T. Gavin
NICR - Centre for Public Health
The Queen's University of Belfast
Mulhouse Building, Grosvenor Road
Belfast BT12 6DP, United Kingdom

doi:10.1016/j.artmed.2016.06.001

Abstract

Background and aims: Machine learning techniques for the text mining of cancer-related clinical documents have not been sufficiently explored. Here some techniques are presented for the pre-processing of free-text breast cancer pathology reports, with the aim of facilitating the extraction of information relevant to cancer staging.

Materials and methods: The first technique was implemented using the freely available software RapidMiner to classify the reports according to their general layout: 'semi-structured' and 'unstructured'. The second technique was developed using the open source language engineering framework GATE and aimed at the prediction of chunks of the report text containing information pertaining to the cancer morphology, the tumour size, its hormone receptor status and the number of positive nodes. The classifiers were trained and tested respectively on sets of 635 and 163 manually classified or annotated reports, from the Northern Ireland Cancer Registry.

Results: The best result of 99.4% accuracy – which included only one semi-structured report predicted as unstructured – was produced by the layout classifier with the *k* nearest algorithm, using the binary term occurrence word vector type with stopword filter and pruning. For chunk recognition, the best results were found using the PAUM algorithm with the same parameters for all cases, except for the prediction of chunks containing cancer morphology. For semi-structured reports the performance ranged from 0.97 to 0.94 and from 0.92 to 0.83 in precision and recall, while for unstructured reports performance ranged from 0.91 to 0.64 and from 0.68 to 0.41 in precision and recall. Poor results were found when the classifier was trained on semi-structured reports but tested on unstructured.

doi:10.1016/j.artmed.2016.06.001

Conclusions: These results show that it is possible and beneficial to predict the layout of reports and that the accuracy of prediction of which segments of a report may contain certain information is sensitive to the report layout and the type of information sought.

Keywords

Natural language processing; information extraction; supervised machine learning; surgical pathology report; cancer staging.

doi:10.1016/j.artmed.2016.06.001

1 Introduction

In a recent review, I. Spasić et al. [1] have analysed published efforts for the text mining of cancer-related information from a range of clinical documents. From their analysis, they conclude that machine learning (ML) techniques should be more widely investigated as an alternative to the commonly used rule-based methods. This paradigm shift, they argue, might help resolve the many difficulties faced by traditional methods in dealing with unconventional or erroneous spelling and grammar of clinical documents. A recent natural language processing (NLP) challenge for clinical records [2] has also shown that, while rule-based systems appear to dominate for clinical information extraction tasks, hybrid systems [3] combining ML algorithms and rule-based engines may outperform them. While rule-based approaches do achieve high levels of performance [4] and are usually adopted by commercial tools (such as AutoCode[1]), their application may be less successful in some contexts [5, 6]. More recent projects and contests have consistently shown that ML is a promising methodology in the context of text mining of clinical narratives [7]. In the cancer domain, this is especially true for both for the detection of reportable cases [8, 9] and the actual extraction of relevant information [10, 11]. Those studies also seem to confirm that ML techniques are best deployed alongside rule-based methodologies.

In this study, the performance of off-the-shelf freely available ML tools and techniques which may be used as an aid to the tasks of information extraction from breast cancer pathology reports was explored. Pathology reports are produced by trained clinicians after the macroscopic and microscopic examination of surgically resected tissue specimens and are considered the most authoritative source of cancer diagnosis information. The underlying idea, here, is that the reports can be classified into different types according to their overall structure and that, depending on their classification, fragments of the reports can be identified. These fragments most likely contain the information to be found. The isolation of relevant fragments may both reduce the complexity and

[1] http://www.aim.on.ca/products/autoCode.jsp (Accessed: 6 August 2008).

improve the performance of subsequent rule-based techniques for the actual extraction of information.[2] Document structure and chunk recognition are active areas of research in several fields of application. However, their application to surgical pathology reports has never been explored. In particular, the value of automatically recognising the structure and relevant sections of reports in the context of information extraction has been neglected.

The remainder of this introduction will provide some background to the problem. The following section 2 will outline the methods for the two main tasks – layout classification and chunk recognition. Section 3 illustrates the various tests and evaluation tasks performed for layout classification (3.1) and chunk recognition (3.2), followed by a discussion (section 4). Finally, a brief summary and some reflections on further research that might follow from this work are provided in section 5. Further details of the software tools used, of the textual features of the corpus and the pre-processing operations that were required on it are collected in the Supplementary Materials, together with a Glossary of medical terms, sample documents and novel source code.

1.1 Background

One of the available routes to reducing years of life lost due to cancer illness, at the point of care, is to ensure that epidemiological research is based upon high quality data on the incidence, prevalence and survival rates of cancer: high quality data on cancer episodes constitutes vital support for epidemiological cancer research, cancer care auditing and assessment [12]. In particular, improved completeness of staging information at diagnosis (i.e. the extent of spread of the cancer) allows for improved treatment planning and assessment of treatment effectiveness [13] and provides data for survival analysis closer to the true outcomes of the diseases [14].

The present study was conducted in the Northern Ireland Cancer Registry (NICR), which has received in electronic format all cancer-related pathology reports from all laboratories in Northern Ireland

[2] In some cases, information extraction precision may increase by up to 19% (11%) for structured (unstructured) reports (publication in preparation).

since 1993.[3] We concentrated on the reports stored in the NICR for breast cancers diagnosed in 2006.[4] Although the cancer registration dataset in the UK has been approved [15] and hopefully a higher level of integration of IT systems within the National Health Service [16] will be achieved, currently the values of many clinical test results, staging information and other pathology-related data items are not captured automatically by the main database system of the registry. Because such information is not recorded at the source (laboratory computer systems), it is not received from the data providers as a specific item of the dataset. As a result, all such data are either not available or have to be obtained by human inspection of the free text pathology reports or by the application of ad-hoc techniques. The aim of the NICR is to achieve 70% completeness in staging information. In other words, 70% of all cancer registrations should eventually be associated with some acceptable form of staging information in the NICR database. This is in alignment with the same target agreed by the UK and Ireland Association of Cancer Registries for all their registries [17]. At the same time, particular emphasis has also been placed on the importance of collecting data on some specific hormone receptor protein expression for breast cancer. In some cases the data provided from the employment of these biomarkers can indicate the effectiveness of particular treatment strategies [18] and can be used to achieve accurate patient stratification and deliver personalised medicine [19]. For these reasons, it was decided to focus the study on breast cancer reporting and on this specific information contained in the pathology reports, in addition to another essential piece of information about cancer, namely its morphology [20]:

- TNM staging [21]. This is the most commonly used cancer staging classification system. It includes information on the extent of the primary tumour (T), the absence or presence of lymph node metastasis (N) and the absence or presence of distant metastasis (M).
- Cancer morphology, which is the type of cancer cells and affects the behaviour of the disease [22].

[3] U.S. cancer registries also have increasing access to full text pathology reports, with several of them achieving > 90% population-based coverage (Eric Durbin, personal communication, June 2015).

[4] From now on these will simply be referred to as “the reports”.

- Hormone receptors (oestrogen, herceptin, progesterone).

1.2 Previous work

Document classification is a vast topic. Automatic document classification systems have already been developed as layers within more complex systems, for example aiming at information filtering [23]. However, in the domain of document *layout* classification, available research is usually devoted to the classification of documents under one of a number of already known, strictly well-defined semi-structured layouts. Tresch et al. [24], for instance, show how vector space classification can be used to determine the type of semi-structured documents (e.g. LaTeX document or XML) on the basis of features of their content. However, we are not aware of research exploring the possibility to detect the presence of structure *per se*.

Similarly, *chunk recognition* has already been used as an intermediate step towards the extraction of information, by isolating the phrases or arbitrary sections of text which may contain the information sought [25]. This, however, has never been applied to surgical pathology reports or in combination with layout classification.

2 Materials and methods

2.1 Pathology report general layout and content

Preliminary analysis showed that the reports could be roughly classified into two main categories, referred to here as *semi-structured* and *unstructured*. Semi-structured reports display several sections, each with a heading comprising one or more paragraphs. These headings originate from document templates that the clerical staff in the laboratories use to write down the dictated reports. These templates, however, can be fully edited while typing, making them unreliable for information

extraction.[5] Each section is homogenous regarding the nature of the information contained, and is usually telegraphic with only some sections containing some narrative explanation. Unstructured reports only present a few headings for the main sections, while the corpus of the report with the most relevant information (e.g. nodal status, tumour size and final diagnosis) is a narrative of variable length. Examples of unstructured and semi-structured reports are shown in the Supplementary Materials.

2.2 Methods

Documents and sentences were represented as vectors in a vector space model [26], each dimension corresponding to a term and measured along a number of metrics: term occurrence (TO, the number of occurrences of the term in the document); binary term occurrence (BTO, set to 1 only if TO>0, set to 0 otherwise); term frequency (TF, given by the TO divided by the total number of terms in the document) and term frequency-inverse document frequency (TF-IDF, given by the TF divided by the frequency of the term in the whole corpus).

Supervised ML [27] was used for both tasks, with 5-fold cross-validation employed for both parameter optimisation (on the training set) and evaluation (on the test set). In this study experiments with three classification algorithms were conducted: *k* nearest neighbours (K-NN), Naïve Bayes classification (NB) and Perceptron Algorithm with Uneven Margins (PAUM). Evaluation was based on the standard metrics of accuracy, recall, precision and F1 [28].

Layout classification, that is the labelling of a report as semi-structured or unstructured, was used to provide the subsequent chunk recognition module with two sets of documents which may benefit from distinct processing techniques. The RapidMiner software [29] was employed for the core ML tasks of layout classification.

[5] In the U.S., while pathology labs follow template guidelines according to the College of American Pathologists, most do not capture that data using electronic checklists, but still capture checklist elements within their narrative reports (Eric Durbin, personal communication, June 2015).

Chunk recognition [30] here refers to the automatic identification of those portions of a report where the information sought is most likely to be found. The main aim of this identification would be to provide a NLP module with selected portions of text of interest, the hypothesis being that the reduction of background noise would allow for more effective extraction of the information. The ML facilities provided by the GATE framework [31] were used for all ML operations of this task: manual annotation, preliminary assessments, training and evaluation.

2.3 Document pre-processing and initial analysis

The initial set of reports was pre-processed by Perl routines, shown in the Supplementary Materials. Perl routines were used to remove spurious lines inserted in the body of the reports by the laboratory information systems, and (on the training set) to extract all text fragments that may be the heading of report sections. This latter process was intended as a method to suggest which laboratories may have produced semi-structured reports, although the classification of reports as either semi-structured or unstructured for the supervised learning was manual. The identification of potential headings was also used to assess whether section names could be used to firmly identify portions of reports mentioning information of interest. From this preliminary analysis it could be concluded that information extraction could not be reliably guided by the literal content of section headings only, not even for semi-structured reports. As a result, a rule-based approach was dismissed and the need for exploring more complex statistical methods to aid the identification of relevant sections of pathology reports was confirmed.

Additional automated tools and procedures were also used to help identify reports referring to non-breast cancers. All reports were eventually manually reviewed; the test set was pre-processed and reviewed only after all training sessions had concluded.

doi:10.1016/j.artmed.2016.06.001

3 Results

3.1 Classification of semi-structured and unstructured pathology reports

3.1.1 Pre-processing and manual classification

As already explained in the previous section, the structure that is found in some reports is the result of the use of predefined, standard document templates in the laboratories. The structure of these templates guides both the pathologist's style of dictation and the secretary's typing of the report. However, it has also been observed that the use of these templates may vary even within a single laboratory, depending on the individual clinician's preferences. Additionally, even when a template is used, consistent structure is not guaranteed: the template can be edited while typing, with alterations ranging from the change or removal of headings to the insertion of narrative paragraphs anywhere in the report. Conversely, and as a source of even more confusion for any potential classifier, some very short unstructured reports may resemble semi-structured reports, an example of which is given in the Supplementary Materials.

As a result, the distinction between semi-structured and unstructured reports is not clear-cut but a matter of judgement. Here, it was decided to classify as 'semi-structured' those reports showing subdivision in several headed sections; with each headed section being homogenous in content and mostly written in very telegraphic style and often containing a single value or an array of values (e.g. a number, the result of a measure, "yes", "absent"). The 'unstructured' reports were considered as those only showing a few headings for the main sections but mainly consisting of a narrative segment, mostly written in full sentences with an acceptable grammatical style. Two annotators (both members of the IT group of the NICR and familiar with the layout variations of the reports) created the gold standard for this task; a task with no conflictual outcomes. The final distribution of the reports is shown in Table 1.

doi:10.1016/j.artmed.2016.06.001

Classification	*No. of reports*	
	Training	Test
Semi-structured	423	112
Unstructured	212	51
Total	*635*	*163*

Table 1 – Number of reports by type

3.1.2 Evaluation

After parameter optimisation on the training set (see the Supplemental Materials for details), the chosen algorithms were trained on the whole training set and tested on the test set of reports. Narrow ranges of parameter values, close to those that had produced best performance at the optimisation stage, were used. The best result of 99.4% accuracy, with only one semi-structured report predicted as unstructured, was produced by the *k*-NN algorithm using the BTO word vector type with stopword filter and pruning. The single report wrongly classified as unstructured was visually inspected and resulted in a very short report with few headings (see the Supplementary Materials). This conciseness probably accounts for its classification among the unstructured reports, which are more often shorter than the semi-structured ones.

3.2 Chunk recognition of pathology reports

3.2.1 Pre-processing and manual annotation

All the reports were manually reviewed and all relevant sections annotated with the labels "Diagnosis", "Size", "Nodes" or "Receptors" (more details in the Supplementary Materials). These values were used, respectively, to label those sections of the reports which contained the information relevant to the morphology of the tumour, its size, the number of nodes found positive (the latter two are required for TNM staging) and the results of ER, PR or HER2 receptor status investigations.

These feature values were also selected in order to experiment with a variety of cases. The number of nodes found positive is expressed in the reports in many different ways. In particular, in semi-structured reports it may be expressed in up to three short lines of text (one per each level of axillary nodes), where it is encoded in short strings of digits, like the following:

```
Level 2110
```

This indicates that 11 axillary nodes from level 2 were examined and none were showing metastatic cells. The same information, however, may be found expressed more clearly as in the following line:

```
Level II: Number 11, Involved 0
```

The receptors status is also variable in the ways it may be expressed, but its format is substantially different from the node positivity, varying from plain English only (e.g. “The tumour tested positive for Estrogen Receptor overexpression”) to numeric/symbolic only (e.g. “ER+” or “ER 8/8”). The size of the tumour is a numerical value which is potentially easy to confuse with other numerical values mentioned in the reports, especially other linear dimensions such as the specimen size(s) or the lesion distance from the resection margins. Finally, the diagnosis (i.e. the tumour morphology) is usually expressed very clearly by mentioning one of a few standard terms; as such, the chunk of text containing it was judged a potentially unproblematic section to be automatically classified and the ML performance on it was meant to be a ’benchmark’ for its performance.

3.2.2 Evaluation

Two types of chunk recognition learning were evaluated: *sentence-based classification*, in which each sentence in a document is classified, and *chunk-based classification*, in which any fragment of text in a document may be assigned a class. An initial set of *k*-fold cross-validation sessions were run with the PAUM algorithm on the training sets, in both the sentence-based and chunk-based configurations. The results are shown in Table 2, where the values of precision and recall for chunk-based learning refer to *lenient* matching, which considers predicted chunks as positive matches even when a manually annotated chunk partially overlaps with it.

doi:10.1016/j.artmed.2016.06.001

		Semi-structured		Unstructured	
		Precision	***Recall***	***Precision***	***Recall***
Sentence based	***Diagnosis***	0.91	0.83	0.71	0.73
	Receptors	0.93	0.94	0.73	0.55
	Nodes	0.93	0.90	0.33	0.33
	Tumour Size	0.92	0.91	0.20	0.23
Chunk based	***Diagnosis***	0.92	0.90	0.79	0.55
	Receptors	0.94	0.91	0.89	0.83
	Nodes	0.97	0.94	0.69	0.42
	Tumour Size	0.98	0.92	0.35	0.11

Table 2 – K-fold cross-validation (lenient) results, produced with default settings to provide an initial assessment of the method capabilities.

Then the same process of inspecting *k*-fold cross-validation results was repeated while varying algorithm and parameters, as shown in the Supplementary Materials. This was performed with the diagnosis and tumour size annotations, because they were expected to provide an insight of how the detection of two different types of chunks (one composed by words only, the other composed by mostly words with some numeric values) would be affected by those changes.

Because the chunk-based classification performed consistently better than sentence-based on the training sets, it was decided to perform the final evaluation sessions on the test sets with chunk-based classification only. The precision and recall lenient results of this final evaluation for all annotations are shown in Table 3, while Table 4 provides the detailed counts or classifications (exact matches, misses, false positive and partially overlapping chunks). With the exception of diagnosis, for all cases, the PAUM algorithm was used with full token features (category, orthography, kind and root) and range 2. For diagnosis, range 4 was used for the classification of unstructured reports and range 1, with no other features than root, was used for the classification of semi-structured reports.

doi:10.1016/j.artmed.2016.06.001

	Semi-structured		Unstructured	
	Precision	***Recall***	***Precision***	***Recall***
Diagnosis	0.94 ±.03	0.85 ±.05	0.91 ±.06	0.61 ±.09
Receptors	0.97 ±.03	0.83 ±.06	0.88 ±.10	0.68 ±.14
Nodes	0.97 ±.02	0.92 ±.04	0.83 ±.14	0.54 ±.18
Tumour size	0.96 ±.04	0.90 ±.06	0.64 ±.23	0.41 ±.23

Table 3 – Final evaluation of chunk-based classification, with 95% confidence intervals. These are lenient results, which consider predicted chunks as positive matches even when a manually annotated chunk partially overlaps with it.

						Strict			Lenient		
		Match	***Missed***	***False positive***	***Overlap***	***Recall***	***Precision***	***F_1***	***Recall***	***Precision***	***F_1***
Semi-structured	***Diagnosis***	155	33	12	26	0.72	0.80	0.76	0.85	0.94	0.89
	Receptors	78	27	4	53	0.49	0.58	0.53	0.83	0.97	0.89
	Nodes	168	15	5	16	0.84	0.89	0.87	0.92	0.97	0.95
	Tumour size	89	10	4	6	0.85	0.90	0.87	0.90	0.96	0.93
Unstructured	***Diagnosis***	33	40	6	29	0.32	0.49	0.39	0.61	0.91	0.73
	Receptors	21	13	4	7	0.51	0.66	0.58	0.68	0.88	0.77
	Nodes	5	13	3	10	0.18	0.28	0.22	0.54	0.83	0.65
	Tumour size	7	10	4	0	0.41	0.64	0.50	0.41	0.64	0.50

Table 4 – Detailed results of final evaluation, with number of chunks that were matching (the predicted chunk of text coincides with the manually annotated chunk), overlapping (the two chunks only partially

overlap), missing (a manually annotated chunk is not predicted) or spurious (a predicted chunk does not correspond to a manual annotation).

Finally, the same evaluations as in the previous paragraph were performed, with the difference that the algorithm was trained on the *semi-structured* reports training set but the classifier was then tested on the *unstructured* reports test set (*cross-training*). The results are shown in Table 5.

					Strict			Lenient		
	Match	***Missed***	***False positive***	***Overlap***	***Recall***	***Precision***	***F_1***	***Recall***	***Precision***	***F_1***
Diagnosis A	6	88	2	8	0.06	0.38	0.1	0.14	0.88	0.2
Diagnosis B	8	85	2	9	0.08	0.42	0.1	0.17	0.89	0.3
Receptors	3	24	5	14	0.07	0.14	0.1	0.41	0.77	0.5
Nodes	2	24	1	2	0.07	0.4	0.1	0.14	0.8	0.2
Tumour size	0	17	0	0	0	1	0	0	1	0

Table 5 – Detailed results of cross-training, where the algorithm was trained on semi-structured reports and applied to unstructured reports. *Diagnosis A* and *B* results were produced using the optimal setting for the ML algorithm previously found for semi-structured and unstructured reports, respectively.

4 Discussion

4.1 Layout classification

The results shown in section 3.1.2 provide evidence that semi-structured and unstructured reports can be successfully discriminated by means of ML techniques, using supervised learning. This result

is not only promising for the information extraction aims of NLP, but it may also be used as a basis for manual processing optimisation. For example, it can be used to provide human coders different types of reports depending on their experience or training level. This application, of course, goes beyond biomedical reporting and could impact beneficially on the human inspection of any corpus which includes documents in a range of layout formats.

4.2 Chunk recognition

The results of this task reached a wide variety of accuracy levels, depending on the chunk type, the ML algorithm and the corresponding settings.

The recognition of chunks was consistently more successful in semi-structured reports, with precision and recall ranging 0.94 to 0.97 and 0.83 to 0.92 respectively, than the attempt on unstructured reports, where precision and recall varied from 0.64 to 0.91 and from 0.41 to 0.68 respectively (results shown in Table 3 and represented in Figure 1). The differences are particularly significant for recall and these results may have been easily expected: chunks with the same type of information tend to be more homogeneous in semi-structured reports, so that their vector representation is less sparse than that of equivalent chunks in unstructured reports and their classification presents fewer ambiguities. In this view, layout classification may be seen as a preliminary feature selection layer for the subsequent ML processing steps. As such, our results suggest its investigation may be worth in other language processing pipelines involving ML modules, beyond the chunk recognition task investigated here.

doi:10.1016/j.artmed.2016.06.001

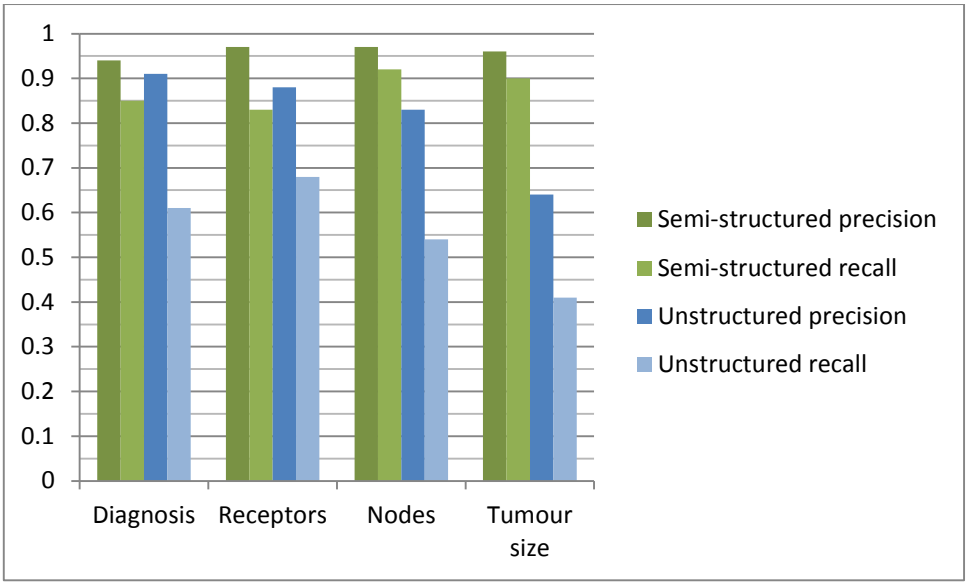

Figure 1 – Graphical representation of the summary results of precision and recall for chunk-based classification, values from Table 3. Chunk recognition is always more successful in semi-structured reports than in unstructured reports.

The results also show that the performance of the chunk recognition in semi-structured reports was less dependent on the type of chunk than in unstructured reports, with the lowest F_1 scores at 94% and 65% of the highest respectively (see Figure 2). In particular, the classification precision of chunks containing diagnosis information was better than all others in unstructured reports, but the recall rate of receptors chunks was better, so that the F_1 score of the classification of receptors chunk was the highest. This was partially anticipated in section 3.2.1.

From the parameter optimisation runs (see the Supplementary Materials) it was also possible to find that the classifier performance on extracting diagnosis chunks was less sensitive to variation of parameters than that on extracting tumour size, in both sentence-based and chunk-based configurations.

doi:10.1016/j.artmed.2016.06.001

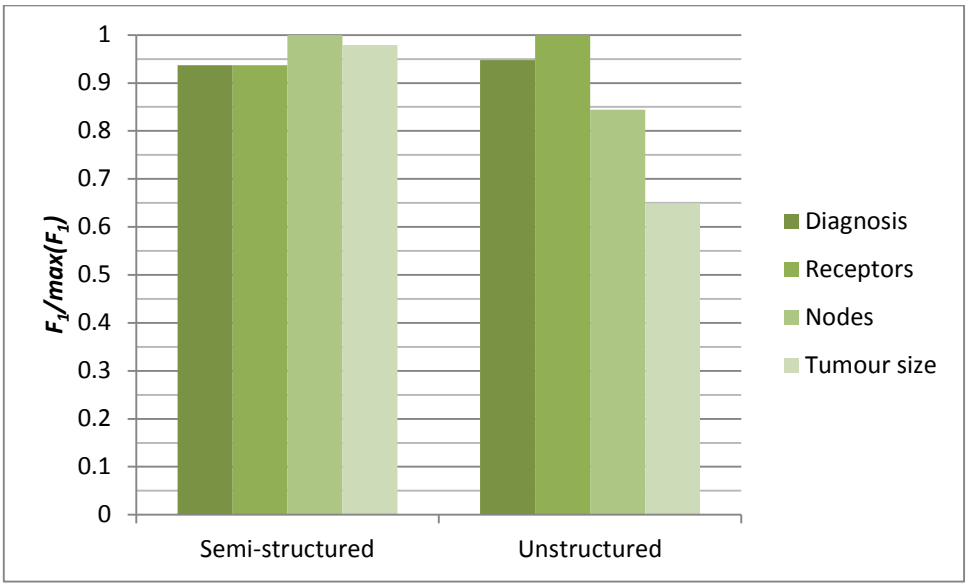

Figure 2 – F_1 score variation in unstructured and semi-structured sets, relative to the top value set to 1 (nodes for semi-structured reports, receptors for unstructured). Performance of chunk recognition in unstructured reports is more variable than in semi-structured reports.

When GATE is used for the application of a trained classifier to a set of annotated test documents, it is possible to examine individual documents and the detail of all the annotations. The examination of sample reports after the final evaluation test indicated that the results summarised in Table 3 and Table 4 should be considered a pessimistic evaluation, due to the subjectivity of the manual annotation process and to some limitations of the GATE mechanism for comparing annotations. In particular there were:

- instances of missing annotations in a report containing a second chunk of the same type, which was correctly predicted;
- instances of spurious predictions (false positives) of chunks which, however, contained relevant information, but had not been manually annotated because more comprehensive information was present elsewhere in the same report;

- cases in which a manually annotated chunk was annotated by the classifier with, for example, three separate chunks. In this instance one prediction was automatically scored as an overlap and the remaining two as spurious because the manually annotated chunk had already been 'consumed' in the partial match with the first predicted chunk.

Additionally, we found that the lenient measures were indeed a good measure of performance, as the non-matching but overlapping chunks only differed by irrelevant portions, such as trailing spaces or full stops. This, again, was due to the slight variability in chunk annotation style even for the same annotator over time.

Finally, the attempt to recognise relevant chunks in unstructured reports by training the algorithm on semi-structured reports (*cross-training*) has shown that this method produces poor results. Although the precision does not fall considerably (see Figure 3), the recall falls by values ranging from 40% to 100%. This implies that it is not possible to use cross-training to 'shortcut' the supervised learning process and reduce the required resources.

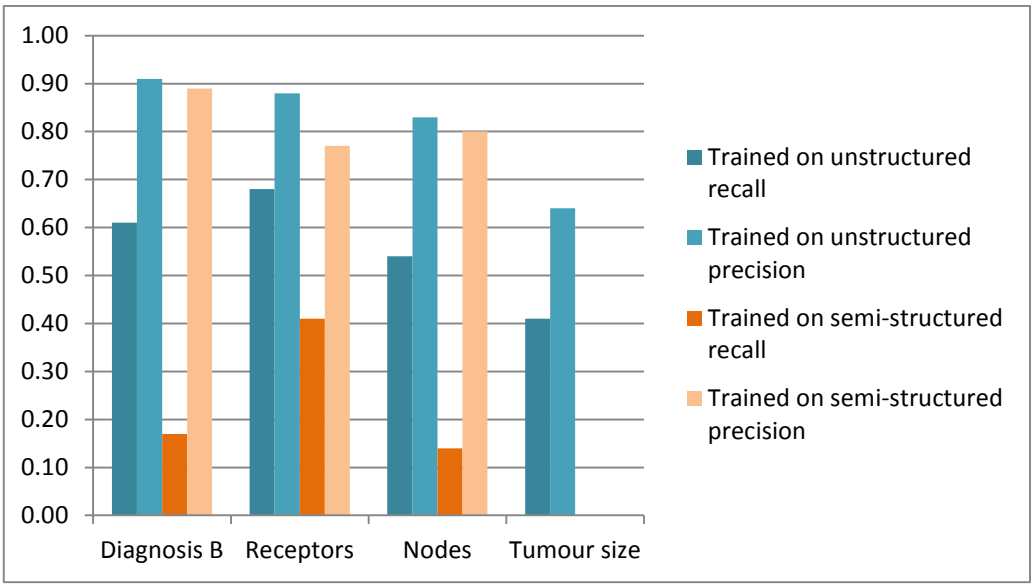

Figure 3 – With the exclusion of tumour size, cross-training did not affect precision dramatically (compare light orange with light blue columns). However, recall was drastically reduced (deep coloured columns): tumour size chunks were not found at all and the recall of receptors chunks was almost halved.

doi:10.1016/j.artmed.2016.06.001

5 Summary and further work

Cancer registries rely on discrete coded data from healthcare systems (such as hospital discharge or pathology databases), which can be in a state of flux over a period of time while clinical coders attempt to determine the most appropriate codes, particularly for diagnosis and staging from source documents. The source documents for cancer registration include radiology, pathology, haematology reports and clinician notes. It is reasonable to suppose that cancer registrations made on the basis of source documents would be faster, more complete and accurate, if 'intelligent' techniques were deployed to process the information contained in such documents.

In this study it was shown that

1. Machine learning can be used successfully to discriminate between semi-structured and unstructured surgical pathology reports.

2. Machine learning can be used with a variable degree of success to identify named or otherwise relevant sections within semi-structured and unstructured reports, depending on the information sought in those sections. Overall, better results are produced with semi-structured reports. Also, it is not possible to train the ML algorithm on semi-structured reports only and apply the resulting chunk classifier to both semi-structured and unstructured reports.

Separate results[6] that we have obtained when combining these ML layers with the final extraction of the actual information of interest, show that information extraction performance may benefit considerably from ML pre-processing (we found up to 19% increase in precision for the extraction of cancer morphology[6]). However this depends on the information sought and the extraction

[6] Publication in preparation.

technique. In particular, while chunk recognition may have high precision in finding fragments that do contain relevant information, even in the lenient case (number of false positives is the same as in the strict case) in some cases it shows low recall. This low recall may interfere with information extraction methods and applications which are less affected by noise (we found 6% decrease in recall for the same cancer morphology extraction scenario[6]). Further work should also be conducted to establish, with more certainty, what features of different types of chunks contribute to their characterisation in terms of word vectors so that more successful classifiers may be designed. Additionally, some tests will be needed to assess the variation of the results found here over time, especially for the purpose of establishing how often the classifiers should be re-trained on newly annotated reports. Finally, applications of these techniques to aid the extraction of information from other healthcare textual sources, such as primary care clinical notes [6], should also be investigated.

6 Acknowledgments

The experimental work for this study was completed at the Northern Ireland Cancer Registry, which is funded by the Public Health Agency NI. We also wish to acknowledge Dr Linda Caughley MBE, honorary pathologist at the NICR, and the staff of the NICR, especially Bernadette Anderson, Colin Fox, Jacky Kelly, Clare Marks, Julie McConnell and Dr Richard Middleton.

7 Conflict of interests

None declared

8 References

[1] Spasić I, Livsey J, Keane JA, Nenadić G, Text mining of cancer-related information: Review of current status and future directions, Int J Med Inf 83 (2014) 605-623.

[2] Uzuner Ö, Solti I, Cadag E, Extracting medication information from clinical text, Journal of the American Medical Informatics Association 17 (2010) 514-518.

[3] Patrick J, Li M, High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge, Journal of the American Medical Informatics Association 17 (2010) 524-527.

[4] Napolitano G, Fox C, Middleton R, Connolly D, Pattern-based text mining of pathology reports for cancer registration, Cancer Causes and Control 21 (2010) 1887-1894.

[5] Johnsi Rani GJ, Gladis D, Manipadam MT, Ishitha G, Breast cancer staging using natural language processing, in: Mauri JL, Thampi SM, Wozniak M, Marques O, Krishnaswamy D, Sahni S et al., eds., Proceedings of the 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI). Kerala (India), IEEE; 2015, August: 1552-1558.

[6] Savkov A, Carroll J, Koeling R, Cassell J, Annotating patient clinical records with syntactic chunks and named entities: the Harvey Corpus, Language Resources and Evaluation (2016) 1-26, first online.

[7] Névéol A, Grouin C, Tannier X, Hamon T, Kelly L, Goeuriot L et al., CLEF eHealth evaluation lab 2015 Task 1b: Clinical named entity recognition, in CLEF 2015 Online Working Notes, CEUR-WS (2015).

[8] Koopman B, Zuccon G, Nguyen A, Bergheim A, Grayson N, Automatic ICD-10 classification of cancers from free-text death certificates, Int J Med Inf 84 (2015) 956-965.

[9] Nguyen DHM, Patrick JD, Supervised machine learning and active learning in classification of radiology reports, Journal of the American Medical Informatics Association 21 (2014) 893-901.

[10] Patrick JD, Nguyen DHM, Wang Y, Li M, A knowledge discovery and reuse pipeline for information extraction in clinical notes, Journal of the American Medical Informatics Association 18 (2011) 574-579.

[11] Martinez D, Li Y, Information extraction from pathology reports in a hospital setting, in: Berendt B, de Vries A, Fan W, Macdonald C, Ounis I, Ruthven I, eds., Proceedings of the 20th ACM International Conference on Information and Knowledge Management. Glasgow, Scotland, UK. New York (NY, USA), ACM; 2011, October: 1877-1882.

[12] Moore GW, Berman JJ, Anatomic pathology data mining, in: Cios KK, ed., Medical Data Mining and Knowledge Discovery (Springer-Verlag, Berlin, 2001).

[13] Davies AR, Deans DAC, Penman I, Plevris JN, Fletcher J, Wall L et al., The multidisciplinary team meeting improves staging accuracy and treatment selection for gastro-esophageal cancer, Diseases of the Esophagus 19 (2006) 496-503.

[14] Walters S, Maringe C, Butler J, Brierley JD, Rachet B, Coleman MP, Comparability of stage data in cancer registries in six countries: Lessons from the international cancer benchmarking partnership, International Journal of Cancer 132 (2013) 676-685.

[15] NHS Information Standards Board, Approved standards: Cancer outcomes and services dataset, standard ISB 1521 (2013), http://www.isb.nhs.uk/library/standard/102 (Accessed: 6 June 2013).

16] NHS connecting for health, http://www.connectingforhealth.nhs.uk (Accessed 10 October 2011).

[17] UKIACR (Ed), Ukiacr Annual Performance Indicators 2014, http://www.ukiacr.org/sites/ukiacr/files/file-uploads/miscellaneous/UKIACR%20Annual%20Report%202014.pdf (Accessed 5 January 2016).

[18] DeVita VT, DeVita Jr VT, Lawrence TS, Rosenberg SA, Cancer: Principles & Practice of Oncology (Wolters Kluwer Health, Philadelphia, 2010).

[19] van't Veer Laura J., Bernards R, Enabling personalized cancer medicine through analysis of gene-expression patterns, Nature 452 (2008) 564-570.

[20] IHTSDO, Systematized nomenclature of medicine, http://www.ihtsdo.org/snomed-ct/ (Accessed: 10 October 2011).

[21] Sobin LH, Wittekind C, UICC TNM Classification of Malignant Tumours, 6th Edition (Wiley-Liss, New York, 2002).

[22] Esteban D, Whelan S, Laudico A, Parkin D, Badger D, Gravestock S et al., eds., Manual for Cancer Registry Personnel (International Agency for Research on Cancer, Lyon, 1995).

[23] Mostafa J, Lam W, Automatic classification using supervised learning in a medical document filtering application, Information Processing & Management 36 (2000) 415-444.

[24] Tresch M, Palmer N, Luniewski A, Type classification of semi-structured documents, in: Dayal U, Gray PMD, Nishio S, eds., Proceedings of the 21st International Conference on Very Large Data Bases. Zurich (Switzerland), Morgan Kaufmann; 1995, September: 263-274.

[25] Appelt DE, Hobbs JR, Bear J, Israel D, Tyson M, FASTUS: A finite-state processor for information extraction from real-world text, in: Bajcsy, ed., Proceedings of the 13th International Joint Conference on Artificial Intelligence. Chambéry (France), Morgan Kaufmann; 1993, August: 1172-1172.

[26] Salton G, Wong A, Yang C, A vector space model for automatic indexing, Commun ACM 18 (1975) 613-620.

[27] Alpaydin E, Introduction to Machine Learning (MIT press, Cambridge MA, 2004).

[28] Van Rijsbergen C, Information Retrieval, 2nd Edition (Butterworths, London, 1979).

[29] Mierswa I, Wurst M, Klinkenberg R, Scholz M, Euler T, YALE: Rapid prototyping for complex data mining tasks, in: Ungar L, Craven M, Gunopulos D, Eliassi-Rad T, eds., KDD '06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia, (PA, USA), ACM; 2006, August: 935-940.

[30] Cunningham H, Maynard D, Bontcheva K, Tablan V, GATE: A framework and graphical development environment for robust NLP tools and applications, in: Isabelle P, ed., Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Stroudsburg (PA, USA), ACL; 2002, July: 168.

[31] Bontcheva K, Tablan V, Maynard D, Cunningham H, Evolving GATE to meet new challenges in language engineering, Natural Language Engineering 10 (2004) 349-373.